0142-0353P-SP

#4
JRB
12-19-01

IN THE U.S. PATENT AND TRADEMARK OFFICE

I N F O R M A T I O N S H E E T

Applicant: BERCHE, Stephane
NAJMAN, Laurent

Application No.:

Filed: March 22, 2001

For: A METHOD OF RECOGNIZING AND INDEXING DOCUMENTS

JC986 U.S. PTO
09/813955
03/22/01

Priority Claimed Under 35 U.S.C. 119 and/or 120:

COUNTRY	DATE	NUMBER
France	03/22/00	2000 03639

Send Correspondence to: BIRCH, STEWART, KOLASCH & BIRCH, LLP
P. O. Box 747
Falls Church, Virginia 22040-0747
(703) 205-8000

The above information is submitted to advise the USPTO of all relevant facts in connection with the present application. A timely executed Declaration in accordance with 37 CFR 1.64 will follow.

Respectfully submitted,

BIRCH, STEWART, KOLASCH & BIRCH, LLP

By Thomas S. Auchterlonie
THOMAS S. AUCHTERLONIE
Reg. No. 37,275
P. O. Box 747
Falls Church, VA 22040-0747

/cqc

(703) 205-8000

IN THE U.S. PATENT AND TRADEMARK OFFICE

Applicant(s): BERCHE, Stephane et al.

Application No.: Group:

Filed: March 22, 2001 Examiner:

For: A METHOD OF RECOGNIZING AND INDEXING DOCUMENTS

L E T T E R

Assistant Commissioner for Patents
Box Patent Application
Washington, D.C. 20231

March 22, 2001
0142-0353P-SP

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55(a), the applicant hereby claims the right of priority based on the following application(s):

Country	Application No.	Filed
FRANCE	2000 03639	03/22/00

A certified copy of the above-noted application(s) is(are) attached hereto.

If necessary, the Commissioner is hereby authorized in this, concurrent, and future replies, to charge payment or credit any overpayment to deposit Account No. 02-2448 for any additional fees required under 37 C.F.R. 1.16 or under 37 C.F.R. 1.17; particularly, extension of time fees.

Respectfully submitted,

BIRCH, STEWART, KOLASCH & BIRCH, LLP

By: Thomas S. Auchterlonie
THOMAS S. AUCHTERLONIE
Reg. No. 37,275
P. O. Box 747
Falls Church, Virginia 22040-0747

Attachment
(703) 205-8000
/cqc

BSKB 703.205-8000
BERCHE et al.
Attorney Docket No. 0142-0353P
18/1

R E P U B L I Q U E F R A N Ç A I S E

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

JC986 U.S. PTO
09/813955
03/22/01

BREVET D'INVENTION

CERTIFICAT D'UTILITÉ - CERTIFICAT D'ADDITION

COPIE OFFICIELLE

Le Directeur général de l'Institut national de la propriété industrielle certifie que le document ci-annexé est la copie certifiée conforme d'une demande de titre de propriété industrielle déposée à l'Institut.

Fait à Paris, le 04 JAN. 2001

Pour le Directeur général de l'Institut
national de la propriété industrielle
Le Chef du Département des brevets

Martine PLANCHE

INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE
SIEGE
26 bis, rue de Saint Petersbourg
75800 PARIS cedex 08
Téléphone : 01 53 04 53 04
Télécopie : 01 42 93 59 30
http://www.inpi.fr

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 01 53 04 53 04 Télécopie : 01 42 94 86 54

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle - Livre VI

cerfa N° 11354*01

REQUÊTE EN DÉLIVRANCE 1/2

Cet imprimé est à remplir lisiblement à l'encre noire DB 540 W / 260899

Réservé à l'INPI

REMISE DES PIÈCES
DATE 22 MARS 2000
LIEU 75 INPI PARIS

N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI 0003639

DATE DE DÉPÔT ATTRIBUÉE PAR L'INPI 22 MARS 2000

Vos références pour ce dossier *(facultatif)* J22053/0044/AD

1 NOM ET ADRESSE DU DEMANDEUR OU DU MANDATAIRE À QUI LA CORRESPONDANCE DOIT ÊTRE ADRESSÉE

CABINET BEAU DE LOMENIE
158, rue de l'Université
75340 PARIS CEDEX

Confirmation d'un dépôt par télécopie ☐ N° attribué par l'INPI à la télécopie

2 NATURE DE LA DEMANDE	Cochez l'une des 4 cases suivantes
Demande de brevet	☒
Demande de certificat d'utilité	☐
Demande divisionnaire	☐
Demande de brevet initiale	N° Date / /
ou demande de certificat d'utilité initiale	N° Date / /
Transformation d'une demande de brevet européen *Demande de brevet initiale*	☐ N° Date / /

3 TITRE DE L'INVENTION (200 caractères ou espaces maximum)

"Procédé de reconnaissance et d'indexation de documents".

4 DÉCLARATION DE PRIORITÉ OU REQUÊTE DU BÉNÉFICE DE LA DATE DE DÉPÔT D'UNE DEMANDE ANTÉRIEURE FRANÇAISE	
	Pays ou organisation Date / / N°
	Pays ou organisation Date / / N°
	Pays ou organisation Date / / N°
	☐ S'il y a d'autres priorités, cochez la case et utilisez l'imprimé «Suite»

5 DEMANDEUR		☐ S'il y a d'autres demandeurs, cochez la case et utilisez l'imprimé «Suite»
Nom ou dénomination sociale		OCE-INDUSTRIES S.A.
Prénoms		
Forme juridique		SOCIETE ANONYME
N° SIREN		
Code APE-NAF		
Adresse	Rue	1, rue Jean Lemoine
	Code postal et ville	94000 CRETEIL
Pays		FRANCE
Nationalité		FRANCAISE
N° de téléphone *(facultatif)*		
N° de télécopie *(facultatif)*		
Adresse électronique *(facultatif)*		

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

REQUÊTE EN DÉLIVRANCE 2/2

Réservé à l'INPI

REMISE DES PIÈCES
DATE
LIEU 22 MARS 2000
75 INPI PARIS
N° D'ENREGISTREMENT
NATIONAL ATTRIBUÉ PAR L'INPI 0003639

DB 540 W /260899

Vos références pour ce dossier : *(facultatif)*		J22053/0044/AD
6 MANDATAIRE		
Nom		
Prénom		
Cabinet ou Société		CABINET BEAU DE LOMENIE
N °de pouvoir permanent et/ou de lien contractuel		
Adresse	Rue	158, rue de l'Université
	Code postal et ville	75340 PARIS CEDEX
N° de téléphone *(facultatif)*		01.44.18.89.00
N° de télécopie *(facultatif)*		01.44.18.04.23
Adresse électronique *(facultatif)*		
7 INVENTEUR (S)		
Les inventeurs sont les demandeurs		☐ Oui ☒ Non **Dans ce cas fournir une désignation d'inventeur(s) séparée**
8 RAPPORT DE RECHERCHE		**Uniquement pour une demande de brevet (y compris division et transformation)**
Établissement immédiat ou établissement différé		☒ ☐
Paiement échelonné de la redevance		**Paiement en trois versements, uniquement pour les personnes physiques** ☐ Oui ☐ Non
9 RÉDUCTION DU TAUX DES REDEVANCES		**Uniquement pour les personnes physiques** ☐ Requise pour la première fois pour cette invention *(joindre un avis de non-imposition)* ☐ Requise antérieurement à ce dépôt *(joindre une copie de la décision d'admission pour cette invention ou indiquer sa référence)* :
Si vous avez utilisé l'imprimé «Suite», indiquez le nombre de pages jointes		

10 SIGNATURE DU DEMANDEUR OU DU MANDATAIRE (Nom et qualité du signataire)

PARIS LE 22 MARS 2000
DAVID Alain
CPI N° 98-0500

VISA DE LA PRÉFECTURE OU DE L'INPI

P. BERNOUIS

La loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses faites à ce formulaire. Elle garantit un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI.

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle - Livre VI

cerfa N° 11235*02

DÉPARTEMENT DES BREVETS

26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 01 53 04 53 04 Télécopie : 01 42 94 86 54

DÉSIGNATION D'INVENTEUR(S) Page N° 1 / 1

(Si le demandeur n'est pas l'inventeur ou l'unique inventeur)

Cet imprimé est à remplir lisiblement à l'encre noire

DB 113 W /260899

Vos références pour ce dossier *(facultatif)*	1J22053-44 FR
N° D'ENREGISTREMENT NATIONAL	00 03639

TITRE DE L'INVENTION (200 caractères ou espaces maximum)

Procédé de reconnaissance et d'indexation de documents

LE(S) DEMANDEUR(S) :

OCE-INDUSTRIES S.A.

DESIGNE(NT) EN TANT QU'INVENTEUR(S) : (Indiquez en haut à droite «Page N° 1/1» S'il y a plus de trois inventeurs, utilisez un formulaire identique et numérotez chaque page en indiquant le nombre total de pages).

Nom		BERCHE
Prénoms		Stéphane
Adresse	Rue	75, Boulevard Saint Germain
	Code postal et ville	75005 PARIS
Société d'appartenance *(facultatif)*		
Nom		NAJMAN
Prénoms		Laurent
Adresse	Rue	20, rue Manin
	Code postal et ville	75019 Paris
Société d'appartenance *(facultatif)*		
Nom		
Prénoms		
Adresse	Rue	
	Code postal et ville	
Société d'appartenance *(facultatif)*		

DATE ET SIGNATURE(S) DU (DES) DEMANDEUR(S) OU DU MANDATAIRE (Nom et qualité du signataire)

Paris, le 5 Juin 2000

A. DAVID
CPI-98 0500

CABINET BEAU DE LOMENIE

Domaine de l'invention

La présente invention se rapporte au domaine spécifique de la reconnaissance et de l'indexation de documents dans une base de données. Elle vise tout particulièrement un dispositif et le procédé correspondant permettant l'indexation de dessins techniques et de formulaires à partir de la reconnaissance de seulement certains de leurs éléments.

Art antérieur

Les procédés de reconnaissance de documents sont multiples et sont bien connus. Ils mettent en œuvre une étape initiale de numérisation suivie d'une étape de segmentation elle même suivie d'une étape de reconnaissance OCR. L'étape de segmentation (découpe du document) peut porter sur tout (cas du « pleine page » classique) ou partie du document.

Toutefois, une telle méthode conventionnelle permettant ensuite une indexation automatique du document n'est envisageable que si le document à reconnaître n'est pas trop complexe. En outre, elle n'est pas appliquée en matière de dessins techniques. En effet, dans ce domaine spécifique, il est procédé seulement à la première étape de numérisation, les étapes de segmentation et de reconnaissance étant remplacées par une étape effectuée directement par un opérateur et consistant en une saisie manuelle des éléments particuliers nécessaires à l'indexation du dessin, au niveau de une ou plusieurs zones de ce dessin (présentes en pratique dans son seul cartouche).

On comprend dès lors que ce traitement devienne vite fastidieux lorsqu'il s'agit d'indexer plus d'une dizaine de dessins techniques éventuellement de types différents (c'est à dire avec des cartouches de formes différentes). Il existe donc actuellement un besoin non satisfait pour

une méthode d'indexation à partir d'une reconnaissance automatique de ces dessins.

Objet et définition de l'invention

L'invention se propose donc de résoudre ce problème de façon simple et économique en mettant en œuvre un procédé de reconnaissance et d'indexation de documents consistant, à partir d'un numériseur relié à un ordinateur, tout d'abord à numériser ces documents puis à désigner au moyen d'un organe de pointage de l'ordinateur un point quelconque P d'au moins une case de ces documents et enfin à reconnaître par reconnaissance OCR les caractères de cette case afin de les stocker dans une première base de données reliée à l'ordinateur pour permettre une indexation des documents ainsi numérisés. L'étape de désignation comprend une étape de recherche et d'identification de la case du document à laquelle appartient ledit point P désigné par l'utilisateur.

Ainsi, avec le procédé de l'invention, la saisie manuelle antérieure se limite à une désignation de zones déterminées (appelées cases) à l'intérieure desquelles une reconnaissance automatique des éléments nécessaires à l'indexation d'un premier document de même type sont disponibles. Les documents à reconnaître et à indexer peuvent être constitués par un ensemble de dessins techniques ou de formulaires de type distincts ou non.

L'étape de recherche et d'identification de ladite case est effectuée en appliquant un algorithme de recherche de forme sur une zone de recherche déterminée entourant ledit point P désigné préalablement par l'utilisateur. Cet algorithme de recherche de forme peut être un algorithme à base de transformée de HOUGH ou un algorithme de projection qui compte le nombre de pixels présents dans chaque ligne verticale ou horizontale de ladite zone de recherche déterminée et, à partir de ces nombres, retrouve par l'examen de pics de profils de projection selon X et Y, les lignes horizontales et verticales présentes dans cette zone de recherche.

Ainsi, en limitant la surface à laquelle l'algorithme est appliqué, on peut augmenter notablement sa vitesse d'exécution tout en limitant le nombre d'itérations nécessaires pour reconnaître une case du document.

De préférence, l'étape de numérisation est effectuée tout d'abord pour l'ensemble des documents à exploiter, lesdites étapes d'identification de la case et de reconnaissance OCR de son contenu étant ensuite effectuées successivement pour chacun de ces documents. Toutefois, cette étape de numérisation peut aussi être effectuée tout d'abord pour un premier document, lesdites étapes d'identification de la case et de reconnaissance OCR de son contenu étant ensuite effectuées pour ce même document, ces trois étapes se répétant successivement pour l'ensemble des documents à exploiter.

L'invention se rapporte également au dispositif de reconnaissance et d'indexation de documents mettant en œuvre le procédé précédent. Avantageusement, ce dispositif comporte en outre une seconde base de données reliée à l'ordinateur pour stocker des données (dites données de caractérisation) permettant une identification ultérieure automatique de la case sans désignation préalable d'un point quelconque P de cette case.

Les bases de données peuvent être intégrées dans la mémoire de l'ordinateur ou extérieures à celui-ci. L'organe de pointage peut être remplacé par le clavier de l'ordinateur ou bien encore le doigt de l'utilisateur.

Brève description des dessins

D'autres caractéristiques et avantages de la présente invention ressortiront mieux de la description suivante, faite à titre indicatif et non limitatif, en regard des dessins annexés, sur lesquels:

- la figure 1 représente un exemple de dessin technique,
- la figure 2 montre un exemple de cartouche d'un dessin conforme à la figure 1,

- la figure 3 illustre la structure matérielle générale du dispositif de reconnaissance et d'indexation de documents selon l'invention,
- la figure 4 est un organigramme explicitant le fonctionnement du dispositif de la figure 3 lors de la reconnaissance et l'indexation du dessin de la figure 1,
- la figure 5 est un organigramme détaillant la fonction de recherche et d'identification d'une case du cartouche de la figure 2,
- la figure 6a représente une première zone de recherche incorporant une partie de la case à identifier,
- les figures 6b et 6c illustrent des profils de projection obtenus à partir de la zone de recherche de la figure 6a,
- la figure 7a représente une seconde zone de recherche incorporant une partie plus importante de la case à identifier,
- les figures 7b et 7c illustrent des profils de projection obtenus à partir de la zone de recherche de la figure 7a,
- la figure 8a représente une troisième zone de recherche entourant complètement la case à identifier, et
- les figures 8b et 8c illustrent des profils de projection obtenus à partir de la zone de recherche de la figure 8.

Description détaillée d'un mode préférentiel de réalisation

Conformément aux figures 1 et 2, un dessin technique tel qu'un plan industriel 10 se compose essentiellement du dessin lui même 12 et d'un cartouche 14 comportant plusieurs cases rectangulaires de dimensions diverses.

Ces différentes cases qui portent pour la plupart d'entre elles des mentions particulières ne présentent pas toutes le même intérêt pour une indexation d'un dessin technique. Il en est ainsi par exemple de la mention de la projection, de l'échelle ou du format de ce dessin. Parmi les mentions qui importent lors d'une indexation, on distingue en général au moins une

première case 16 comportant un numéro d'identification du dessin, une deuxième case 18 comportant un titre du dessin et une troisième case 20 comportant une mention de l'auteur du dessin. Bien entendu, ces trois mentions ne doivent en aucun cas être considérées comme limitatives, et on pourrait aussi bien envisager de prendre également en compte une date de la dernière mise à jour accessible dans une quatrième case 22 ou un numéro de planche repérable dans une cinquième case 24.

La figure 3 montre l'architecture matérielle minimale nécessaire à un ensemble informatique pour permettre, selon l'invention, la reconnaissance et l'indexation de documents du type de la figure 1.

Cet ensemble comporte tout d'abord un numériseur ou scanner 30 pour effectuer une numérisation de documents (en l'espèce des plans) devant ensuite être indexés. Ce numériseur est relié à un ordinateur ou micro-ordinateur de type conventionnel 32 muni de moyens logiciels 100 connus pour assurer cette numérisation. Une première base de données 34 reliée également à l'ordinateur 32 est prévue pour stocker les documents ainsi numérisés. On notera, que selon la capacité de stockage interne de cet ordinateur et le volume des données correspondant aux documents à numériser, cette première base de données 34 peut être soit externe, comme illustré, soit directement logée en interne dans l'ordinateur. L'ordinateur comporte bien entendu des moyens logiciels 110 de gestion (création, consultation, modification) de cette première base.

Pour assurer l'indexation des documents au niveau de la première base de données 34, il est prévu que l'ordinateur 32 comporte également des moyens logiciels 120 de reconnaissance OCR de type connu pour reconnaître et identifier certains éléments particuliers de ces documents. Toutefois, ces moyens de reconnaissance OCR sont commandés sous l'action de moyens logiciels spécifiques 130 en liaison avec une seconde base de données 38 contenant des données de caractérisation et permettant un traitement particulièrement simple et rapide de ces documents.

En effet, selon l'invention, cette reconnaissance est effectuée seulement dans des zones déterminées du document, plus particulièrement, dans le cas d'un dessin technique, dans des cases de son cartouche localisées par l'utilisateur au moyen d'un organe de pointage 36 de l'ordinateur, tel qu'une souris, une boule de pointage ou tout autre dispositif équivalent (y compris le doigt de l'utilisateur dans le cas de recours à un écran tactile), lequel permet la désignation d'un point quelconque P de cette case. Eventuellement, en complément, pour améliorer encore le traitement, ces moyens logiciels 130 peuvent proposer à l'utilisateur à l'issue de cette opération de désignation de définir le type de données à reconnaître dans la case ainsi désignée, par exemple une suite de caractères numériques (pour le numéro d'identification) ou une suite de caractères alphanumériques (pour le titre ou le nom de l'auteur par exemple).

Le procédé mis en œuvre dans le dispositif précédent, illustré à la figure 4, suit ainsi les étapes suivantes. Après une numérisation d'un premier document dans une première étape 200 par le numériseur 30 associé aux moyens logiciels 100, il est procédé dans une deuxième étape 210 à un stockage intermédiaire de l'image de ce document au niveau de la mémoire de l'ordinateur 32 ainsi, éventuellement simultanément, qu'à son affichage sur l'écran de visualisation de l'ordinateur (après si nécessaire une opération d'agrandissement dite aussi de « zoom »). Si les moyens logiciels de traitement 130 ne peuvent identifier le type de document numérisé à partir des données issues de la base de données de caractérisation 38 (test de l'étape 220), il est alors procédé à cette identification au cours des étapes suivantes du processus, et notamment, il est tout d'abord opéré, dans une étape 230, au moyen de l'organe de pointage 36 associé à ces moyens logiciels 130, à une désignation par l'utilisateur d'un point P d'une première zone déterminée de ce document, par exemple la case 16 du cartouche 14 du dessin comportant le numéro d'identification de ce dessin. Eventuellement, de façon facultative, comme l'illustre en pointillé l'étape 240, il est possible

que l'utilisateur précise alors le type de caractères qui devront être reconnus dans cette case. Cette indication permet de limiter le choix des caractères à reconnaître (par exemple les seuls caractères numériques 0 à 9) et donc d'améliorer l'étape de reconnaissance OCR ultérieure. A partir de la désignation de ce point (dont les coordonnées sont alors déduites par rapport à un point origine prédéterminé), il est procédé dans une nouvelle étape 250 à la recherche et l'identification de la case à laquelle appartient ce point P (c'est à dire à celle de ou des lignes frontières de cette case comme explicité plus avant en regard de la figure 5) et, une fois cette identification effectuée (par exemple en affichant en surbrillance ou en couleur les contours de cette case) et ses éléments de caractérisation stockés dans la seconde base de données 38 dans une étape 260 (les coordonnées dimensionnelles de la case et la position de son centre sont ainsi mémorisées), il est procédé classiquement dans l'étape immédiatement suivante 270 à la reconnaissance OCR des caractères de cette case grâce aux moyens logiciels connus 120, la fin de cette opération de reconnaissance étant matérialisée par exemple par le fait que l'ordinateur « rend la main » à l'utilisateur.

Les cinq étapes précédentes 230, éventuellement 240, 250, 260 et 270 sont ensuite reprises pour une seconde zone déterminée, puis une suivante, jusqu'à une complète identification du document, c'est à dire jusqu'à ce que toutes les zones nécessaires à son indexation, et déterminées préalablement selon l'utilisation souhaitée, au niveau des moyens logiciels 110, soient prises en compte. Une fois cette opération effectuée, il est procédé, dans une nouvelle étape 280, au stockage de l'image numérisée dans la première base de données 34. Toutes les étapes précédentes sont répétées éventuellement pour un second type de document et, ainsi de suite, jusqu'à épuisement des documents à numériser et indexer. La consultation de la première base 34 sera ensuite possible par les moyens logiciels 110 qui permettront classiquement d'accéder à chacun des documents de la base

selon le critère choisi par l'utilisateur et correspondant à un ou plusieurs des éléments d'indexation retenus initialement.

En effet, et ceci est très important, les opérations de désignation précédentes ne sont réalisées que lors de l'indexation d'un premier document d'un type donné car, si les documents suivants à exploiter sont de même type, il est alors répondu par l'affirmative au test de l'étape 220 et un pointage des mêmes différentes zones supports de l'indexation n'est alors plus nécessaire. Les moyens logiciels 130 ayant mémorisés les coordonnées des cases reconnues à l'issue des premières désignations dans la base de données de caractérisation 38, il leur suffit alors simplement de rechercher à partir du point origine ces mêmes cases dans les documents suivants (cette ressemblance est testée sur la surface de la case et avec une certaine tolérance comme expliqué en regard de la figure 5) et après leur identification d'en analyser automatiquement le contenu par la reconnaissance OCR, sans la désignation préalable d'un point quelconque de ces cases.

On comprend dès lors aisément que le procédé de l'invention est particulièrement rapide et efficace, puisque pour un ensemble de documents semblables, une fois la première identification d'un type donné de document, au cours de laquelle l'intervention de l'utilisateur est indispensable, les suivantes peuvent se poursuivre automatiquement sans nouvelle action de cet utilisateur. A chaque fois, la reconnaissance OCR ne porte que sur les éléments indispensables à l'indexation des documents et non sur l'ensemble de celui-ci, comme dans l'art antérieur.

On notera également qu'à la numérisation « à l'unité » précitée (un document après l'autre), il est possible de substituer une numérisation par lot ou bien encore une numérisation complète (et alors automatique) de l'ensemble des documents à traiter (et à un stockage correspondant dans l'ordinateur), les étapes d'identification et de reconnaissance OCR s'effectuant seulement ensuite successivement pour chaque document de cet

ensemble, une fois cette opération initiale de numérisation entièrement réalisée.

La figure 5 montre les différentes opérations réalisées par le sous programme de recherche mis en œuvre dans les moyens logiciels 130 et destiné à identifier une case déterminée à partir de la seule désignation par l'utilisateur d'un point P de cette case. Ces opérations sont basées sur l'application d'un algorithme de recherche de forme tel qu'un algorithme de projection ou une transformée de HOUGH (pour les formes rondes). En l'espèce, il est procédé à une application particulière d'un algorithme de projection connu en soi et qui consiste à compter le nombre de pixels présents dans chaque ligne verticale ou horizontale d'une image et, à partir de ces nombres, de retrouver par des profils de projection selon X et Y, les lignes horizontales et verticales de cette image (qui sont déterminées par des pics dans ces profils de projection). Cet algorithme présente l'intérêt de procurer un rapport signal/bruit très élevé, car un éventuel « trou » dans une ligne (l'absence d'un pixel) modifie peu la hauteur d'un pic, de même qu'une éventuelle inclinaison d'une ligne n'affecte que peu la position de ce pic.

Toutefois, selon l'invention, cet algorithme de projection n'est pas appliqué à l'ensemble du document mais simplement à une zone déterminée de celui-ci (d'aire Si définie dans une étape première 300) définie autour du point désigné P lors de l'étape de pointage 220. Ainsi, à supposer que cette zone de recherche comprend entièrement la case à reconnaître, il suffit alors seulement d'effectuer une projection de toutes les lignes verticales à droite du point P pour retrouver le coté droit de la case (ce sera celle dont le pic est le plus important ou supérieur à un seuil donné). On fera de même avec les lignes verticales à gauche de ce point pour le coté gauche de la case et avec les lignes horizontales en haut et en bas de ce point pour retrouver respectivement les cotés haut et bas de cette case. Toutefois, en pratique, cette zone de recherche initiale est soit comprise dans celle de la case à

identifier soit à cheval sur celle-ci (voir par exemple l'aire S1 de la figure 6a), et il convient donc d'accroître sa surface progressivement (par paliers déterminés successifs) jusqu'à ce qu'elle comprenne entièrement cette case pour parvenir à cette identification (voir l'aire S3 de la figure 8a). A chaque fois, il est fait application de l'algorithme de projection (étape 310). L'identification est achevée (test de l'étape 330) lorsque pour deux aires successives les positions des pics de projection déterminées à l'étape précédente 320 restent invariables. Les coordonnées de la case trouvée sont alors mémorisées dans une étape suivante 340 pour pouvoir ensuite être utilisées pour une reconnaissance automatique des documents suivants. Un exemple de mise en œuvre de l'algorithme est illustré en regard des figures 6a à 8c qui montrent le processus mis en œuvre pour l'identification par exemple de la case 18 contenant une information à indexer.

On supposera que l'utilisateur a « cliqué » à l'extrême droite de cette case. Les moyens logiciels 130 créent alors une première zone de recherche rectangulaire d'aire S1 autour de ce point qui, comme l'illustre la figure 6a, va comprendre un coté vertical droit 400 et deux parties des cotés horizontaux haut 402 et bas 404 de la case à identifier. L'application de l'algorithme de projection à cette première zone de recherche conduit aux projections horizontales et verticales des figures 6b et 6c. On remarque très bien, sur la figure 6b, les deux pics 412, 414 correspondant aux cotés horizontaux respectifs 402, 404, comme sur la figure 6c, on peut noter le seul pic 410 correspondant au coté vertical 400. Cette première analyse ne permettant pas l'identification de la case 18, il est procédé à un examen automatique d'une deuxième zone de recherche d'aire S2 qui, comme le montre la figure 7a, intègre toujours le coté vertical droit 400 et une partie, toutefois plus importante, des deux cotés horizontaux 402, 404. Le résultats des algorithmes de projection horizontale et verticale sont donnés aux figures 7b et 7c. On reconnaît les pics 410, 412, 414 et d'autres pics 418, 420, plus ou moins nets et correspondant à la mention « gauche »,

apparaissent à la fois sur la projection horizontale et sur la projection verticale. Enfin, cette seconde application de l'algorithme ne permettant toujours pas une identification complète de la case 18, il est définit automatiquement une troisième zone de recherche d'aire S3 qui cette fois englobe totalement la case 18 (voir la figure 8a), notamment entièrement ses cotés horizontaux 402, 404 mais également son coté vertical gauche 406. La projection horizontale résultant de l'algorithme correspondant est illustrée à la figure 8b avec ses deux pics 412, 414 correspondant aux deux cotés horizontaux 402, 404. Par contre, la projection verticale fait maintenant apparaître, outre la série de pics 420, non seulement le pic 410 correspondant au coté droit 400 de la case 18 mais également un nouveau pic 416 correspondant au coté gauche 406 de cette case, permettant ainsi une parfaite identification de la case 18.

Il est important de noter que, si le procédé et le dispositif de l'invention ont été décrits essentiellement au regard de la reconnaissance et l'indexation de dessins techniques, il est bien entendu envisageable de mettre en œuvre ce procédé pour d'autres types de documents et, notamment, une application particulièrement intéressante est celle de la reconnaissance et l'indexation de formulaires, par exemple de type bon de commande (en matière de vente par correspondance notamment) ou encore feuille d'opérations. En effet, le traitement de tels formulaires suppose actuellement de les caractériser préalablement au moyen de symboles particuliers disposés en des endroits spécifiques de ces formulaires, lesquels symboles permettront ensuite une identification automatique du type de formulaire. Dès lors, la caractérisation d'un formulaire est un processus long et complexe qui ne peut se justifier que pour la numérisation de quantité importante de documents semblables.

Avec la présente invention, cette phase de caractérisation préalable disparaît au profit de l'étape de désignation/identification des seules cases du formulaire à traiter.

Ainsi, le procédé d'identification est particulièrement rapide (ce qui est important quant il ne s'agit de traiter que quelques dessins techniques ou formulaires), simple et utilisable par tout opérateur même très peu qualifié. En outre, il est stable vis à vis de bruits de saisie éventuels résultant du déplacement des documents numérisés.

REVENDICATIONS

1. Procédé de reconnaissance et d'indexation de documents (10) consistant, à partir d'un numériseur (30) relié à un ordinateur (32), tout d'abord à numériser (200) ces documents puis à désigner (250) au moyen d'un organe de pointage (36) de l'ordinateur un point quelconque P d'au moins une case (16-24) de ces documents et enfin à reconnaître par reconnaissance OCR (270) les caractères de cette case afin de les stocker (280) dans une première base de données (34) reliée à l'ordinateur pour permettre une indexation des dessins ainsi numérisés.

2. Procédé selon la revendication 1, caractérisé en ce que ladite étape de désignation comprend une étape de recherche et d'identification de la case du document à laquelle appartient ledit point P désigné par l'utilisateur.

3. Procédé selon la revendication 2, caractérisé en ce que ladite étape de recherche et d'identification de ladite case est effectuée en appliquant un algorithme de recherche de forme sur une zone de recherche déterminée entourant ledit point P désigné préalablement par l'utilisateur.

4. Procédé selon la revendication 3, caractérisé en ce que ledit algorithme de recherche de forme est un algorithme de projection qui compte le nombre de pixels présents dans chaque ligne verticale ou horizontale de ladite zone de recherche déterminée et, à partir de ces nombres, retrouve par l'examen de pics de profils de projection selon X et Y, les lignes horizontales et verticales présentes dans cette zone de recherche.

5. Procédé selon la revendication 3, caractérisé en ce que ledit algorithme de recherche de forme est un algorithme à base de transformée de HOUGH.

6. Procédé selon la revendication 1, caractérisé en ce que ladite étape de reconnaissance OCR est précédée par une étape (260) de définition par l'utilisateur du type de caractère à reconnaître dans ladite case du document.

7. Procédé selon la revendication 1, caractérisé en ce que ladite étape de numérisation est effectuée tout d'abord pour l'ensemble des documents à exploiter, lesdites étapes d'identification de la case et de reconnaissance OCR de son contenu étant ensuite effectuées successivement pour chacun de ces documents.

8. Procédé selon la revendication 1, caractérisé en ce que ladite étape de numérisation est effectuée tout d'abord pour un premier document, lesdites étapes d'identification de la case et de reconnaissance OCR de son contenu étant ensuite effectuées pour ce même document, ces trois étapes se répétant successivement pour l'ensemble des documents à exploiter.

9. Procédé selon l'une quelconque des revendications 1 à 8, caractérisé en ce que lesdits documents à reconnaître et à indexer sont constitués par un ensemble de dessins techniques de type distincts ou non.

10. Procédé selon l'une quelconque des revendications 1 à 8, caractérisé en ce que lesdits documents à reconnaître et à indexer sont constitués par un ensemble de formulaires de type distincts ou non.

11. Dispositif de reconnaissance et d'indexation de documents (10) comportant un numériseur (30) pour numériser un document et délivrer une image de ce document, un ordinateur (32) relié au numériseur pour recevoir cette image numérisée, et une première base de données (34) reliée à cet ordinateur pour stocker cette image numérisée, caractérisé en ce qu'il comporte en outre des moyens logiciels (120, 130) pour désigner, au moyen d'un organe de pointage (36) de l'ordinateur, un point quelconque P d'au moins une case (16-24) de cette image, pour rechercher et identifier la case à laquelle appartient ledit point P désigné par l'utilisateur et pour reconnaître, par reconnaissance OCR, les caractères de cette case afin de permettre une indexation des images ainsi numérisées.

12. Dispositif selon la revendication 11, caractérisé en ce qu'il comporte en outre une seconde base de données (38) reliée à l'ordinateur (32) pour stocker des données (dites données de caractérisation) permettant

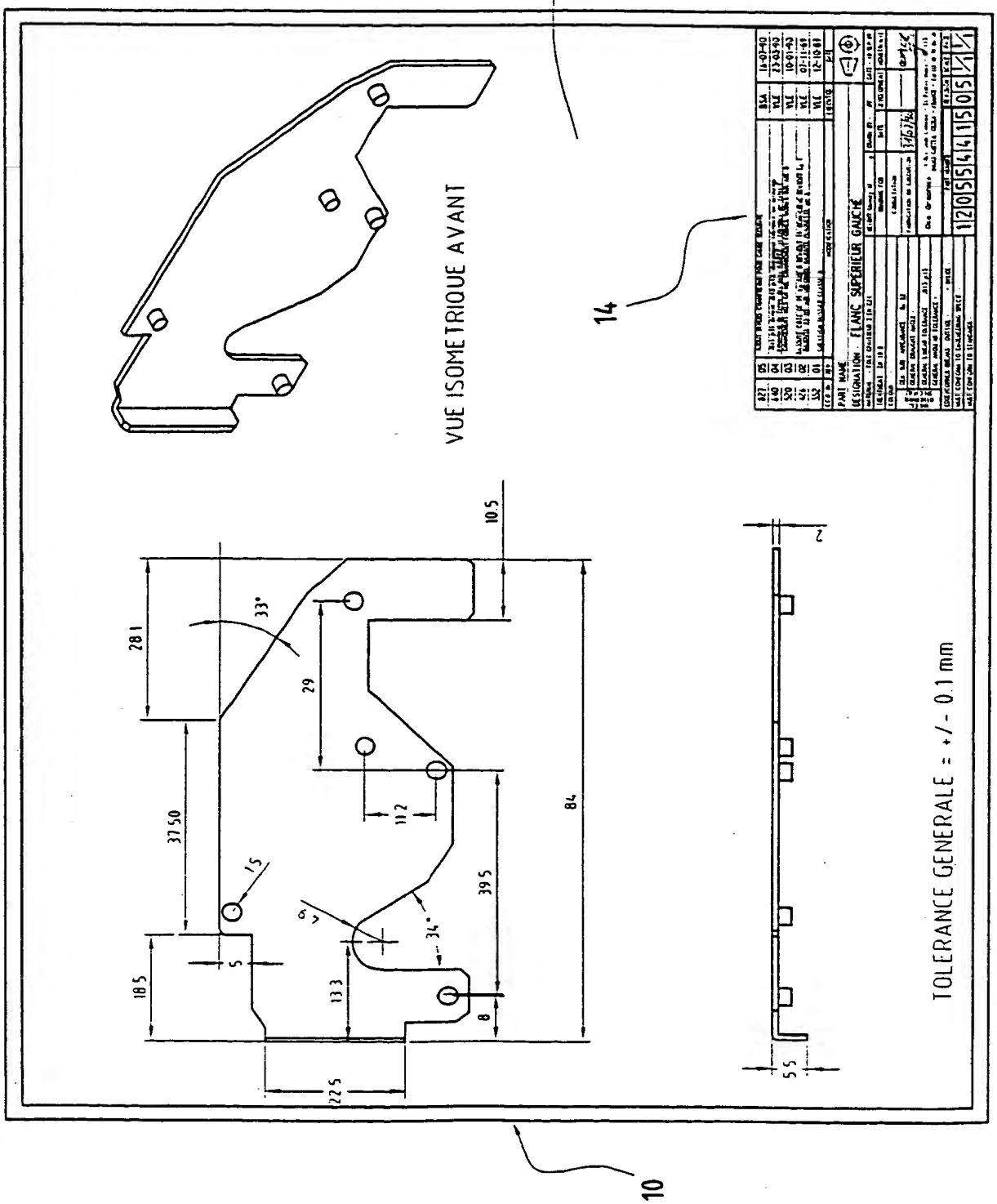
une identification ultérieure automatique de la case sans désignation préalable d'un point quelconque P de cette case.

13. Dispositif selon la revendication 11, caractérisé en ce qu'il comporte en outre des moyens logiciels (120, 130) pour définir le type de données à reconnaître dans ladite case du document.

14. Dispositif selon la revendication 11 ou la revendication 12, caractérisé en ce que les première (34) et deuxième (38) bases de données sont intégrées dans la mémoire de l'ordinateur (32).

15. Dispositif selon la revendication 11, caractérisé en ce que ledit organe de pointage est remplacé par le clavier de l'ordinateur (32) ou le doigt de l'utilisateur.

FIG.1


12
14
10
VUE ISOMETRIQUE AVANT
TOLERANCE GENERALE = +/- 0.1 mm
28.1
33°
29
10.5
84
37.50
11.2
39.5
34°
18.5
5
13.3
8
22.5
15
5.5
2

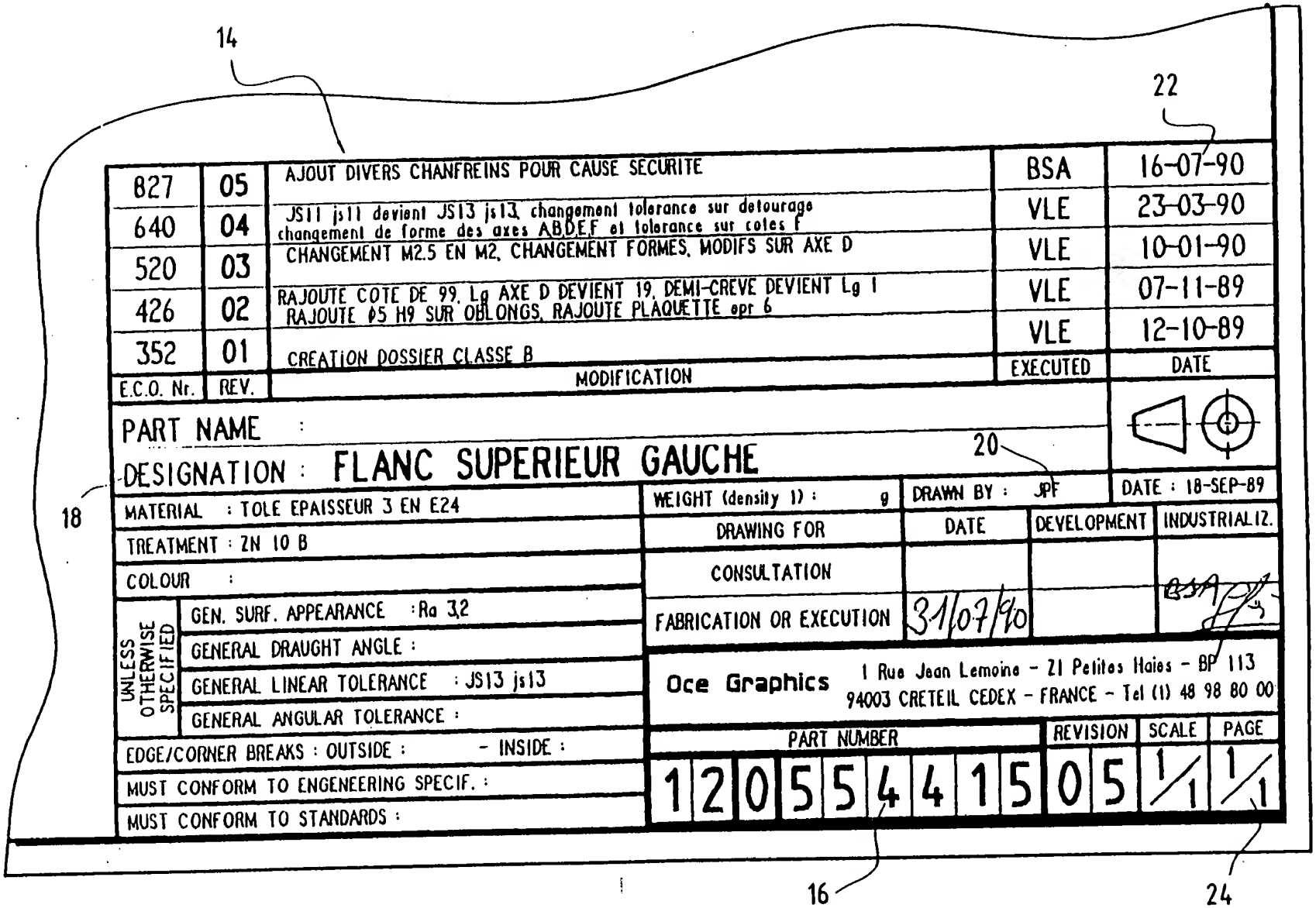

14
22
827 | 05 | AJOUT DIVERS CHANFREINS POUR CAUSE SECURITE | BSA | 16-07-90
640 | 04 | JS11 js11 devient JS13 js13, changement tolerance sur detourage changement de forme des axes ABDEF et tolerance sur cotes f | VLE | 23-03-90
520 | 03 | CHANGEMENT M2.5 EN M2, CHANGEMENT FORMES, MODIFS SUR AXE D | VLE | 10-01-90
426 | 02 | RAJOUTE COTE DE 99, Lg AXE D DEVIENT 19, DEMI-CREVE DEVIENT Lg 1 RAJOUTE ⌀5 H9 SUR OBLONGS, RAJOUTE PLAQUETTE epr 6 | VLE | 07-11-89
352 | 01 | CREATION DOSSIER CLASSE B | VLE | 12-10-89
E.C.O. Nr. | REV. | MODIFICATION | EXECUTED | DATE
PART NAME :
DESIGNATION : FLANC SUPERIEUR GAUCHE
20
18
MATERIAL : TOLE EPAISSEUR 3 EN E24
TREATMENT : ZN 10 B
COLOUR :
UNLESS OTHERWISE SPECIFIED
GEN. SURF. APPEARANCE : Ra 3,2
GENERAL DRAUGHT ANGLE :
GENERAL LINEAR TOLERANCE : JS13 js13
GENERAL ANGULAR TOLERANCE :
EDGE/CORNER BREAKS : OUTSIDE : - INSIDE :
MUST CONFORM TO ENGENEERING SPECIF. :
MUST CONFORM TO STANDARDS :
WEIGHT (density 1) : g
DRAWN BY : JPF
DATE : 18-SEP-89
DRAWING FOR | DATE | DEVELOPMENT | INDUSTRIALIZ.
CONSULTATION
FABRICATION OR EXECUTION 31/07/90
Oce Graphics 1 Rue Jean Lemoine - ZI Petites Haies - BP 113 94003 CRETEIL CEDEX - FRANCE - Tel (1) 48 98 80 00
PART NUMBER | REVISION | SCALE | PAGE
1 2 0 5 5 4 4 1 5 | 0 5 | 1/1 | 1/1
16
24

FIG.2

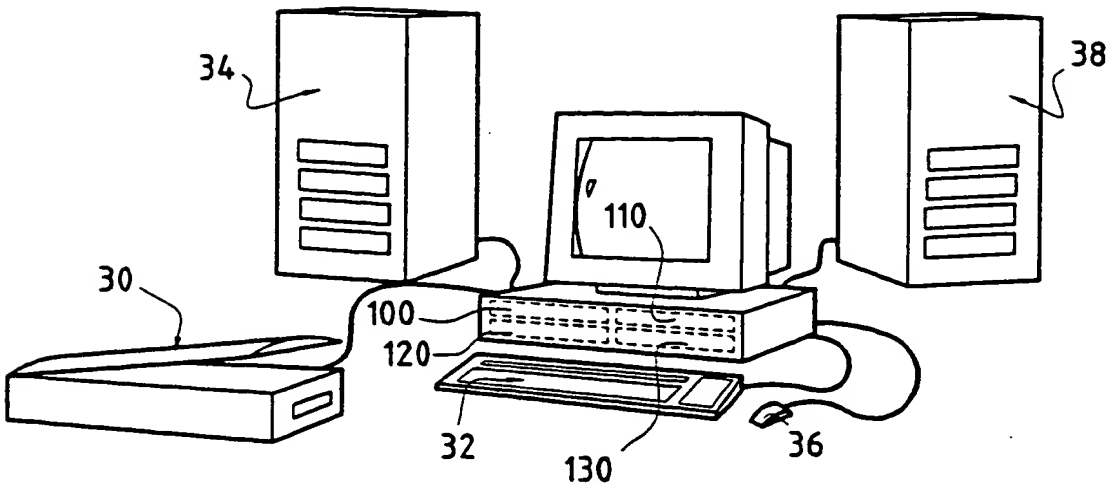

34
38
110
30
100
120
32
130
36

FIG.3

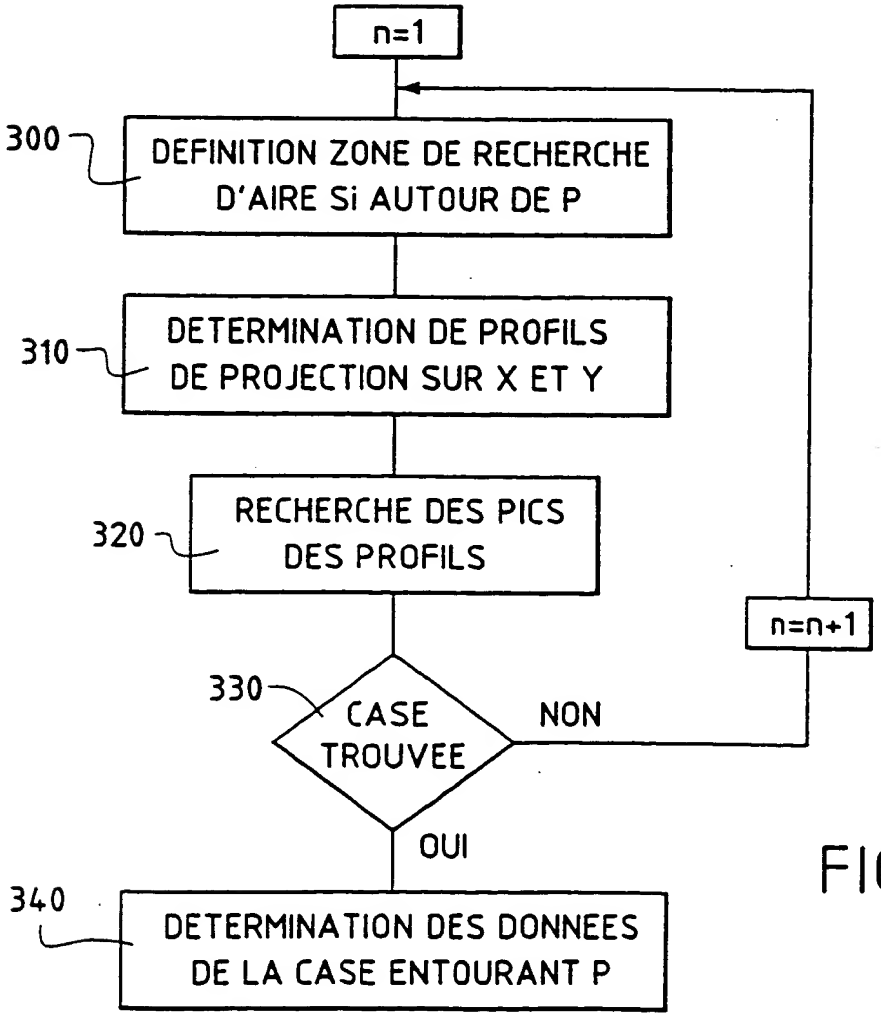

n=1
300
DEFINITION ZONE DE RECHERCHE D'AIRE Si AUTOUR DE P
310
DETERMINATION DE PROFILS DE PROJECTION SUR X ET Y
320
RECHERCHE DES PICS DES PROFILS
n=n+1
330
CASE TROUVEE
NON
OUI
340
DETERMINATION DES DONNEES DE LA CASE ENTOURANT P

FIG.5

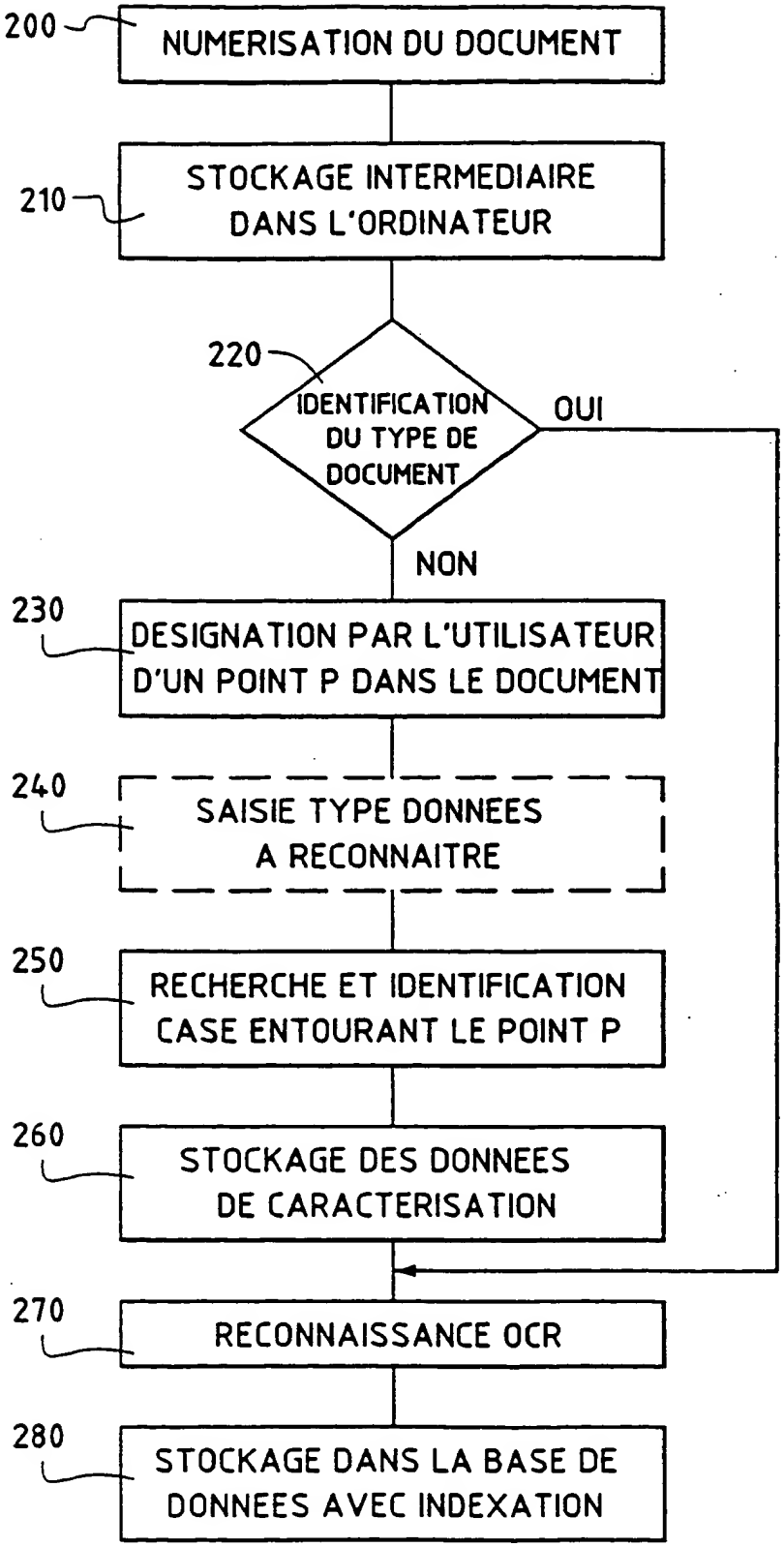

200
NUMERISATION DU DOCUMENT
210
STOCKAGE INTERMEDIAIRE DANS L'ORDINATEUR
220
IDENTIFICATION DU TYPE DE DOCUMENT
OUI
NON
230
DESIGNATION PAR L'UTILISATEUR D'UN POINT P DANS LE DOCUMENT
240
SAISIE TYPE DONNEES A RECONNAITRE
250
RECHERCHE ET IDENTIFICATION CASE ENTOURANT LE POINT P
260
STOCKAGE DES DONNEES DE CARACTERISATION
270
RECONNAISSANCE OCR
280
STOCKAGE DANS LA BASE DE DONNEES AVEC INDEXATION

FIG.4

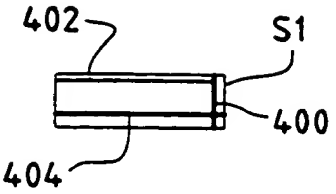


FIG.6A

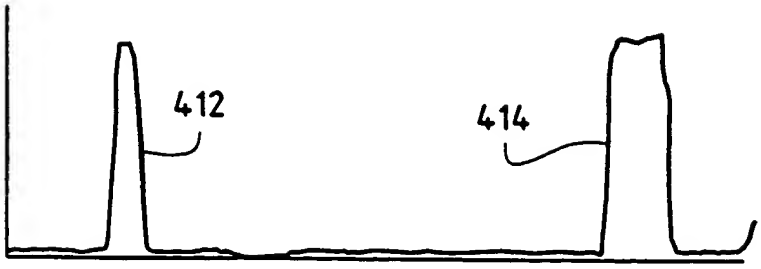


FIG.6B

FIG.6C

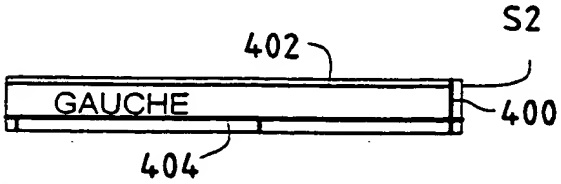


FIG.7A

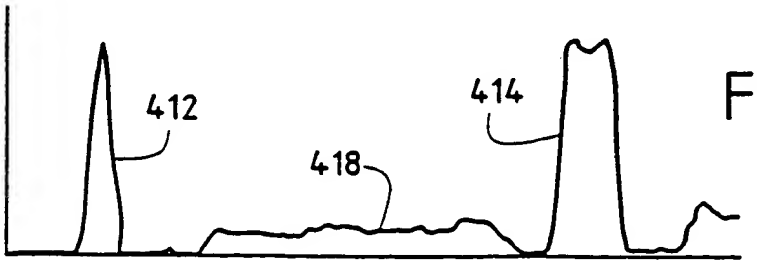


FIG.7B

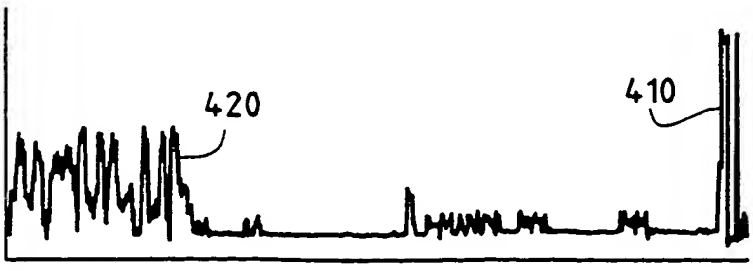


FIG.7C

FIG.8A

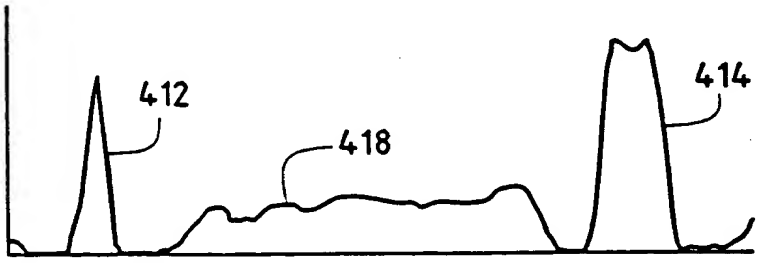


FIG.8B

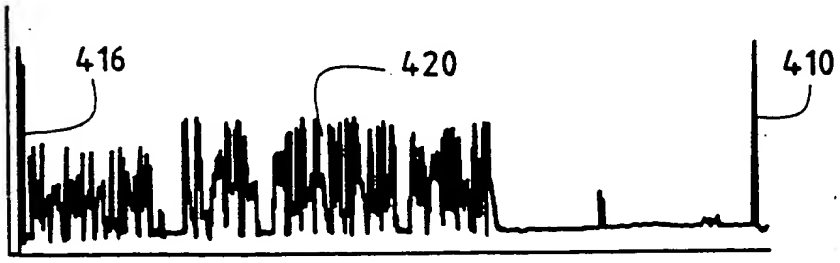


FIG.8C